

A-priori Upper Bounds for the Set Covering Problem

Giovanni Felici¹, Sokol Ndreca², Aldo Procacci³ and Benedetto Scoppola⁴

¹ Istituto di Analisi dei Sistemi ed Informatica, Consiglio Nazionale delle Ricerche, 00185 Roma

²Dep. Estatística-ICEx, UFMG, CP 702 Belo Horizonte - MG, 30161-970 Brazil

³Dep. Matemática-ICEx, UFMG, CP 702 Belo Horizonte - MG, 30161-970 Brazil

⁴Dipartimento di Matematica - Università Tor Vergata di Roma, 00133 Roma, Italy

emails: giovanni.felici@iasi.cnr.it; sokol@est.ufmg.br;

aldo@mat.ufmg.br; scoppola@mat.uniroma2.it

Abstract

In this paper we present a new bound obtained with the probabilistic method for the solution of the Set Covering problem with unit costs. The bound is valid for problems of fixed dimension, thus extending previous similar asymptotic results, and it depends only on the number of rows of the coefficient matrix and the row densities. We also consider the particular case of matrices that are *almost* block decomposable, and show how the bound may improve according to the particular decomposition adopted. Such final result may provide interesting indications for comparing different matrix decomposition strategies.

1 Introduction

Given a finite ground set of objects G and a finite collection \mathcal{G} of its subset, a *Set Cover* \mathcal{C} is a subset of \mathcal{G} such that each element of G is contained in at least one of the subsets in \mathcal{C} . The *Set Covering Problem* (SCP) consists in finding the set \mathcal{C} of minimum cardinality. If positive weights are attached to each element of \mathcal{G} , the *weighted* version of SCP consists in finding a set \mathcal{C} for which the sum of the weights of its element is minimum. Non weighted SCP may also be referred to as SCP with unit costs. If m and n denote the cardinalities of G and \mathcal{G} respectively, the SCP is usually reformulated as the problem of covering the rows of a $m \times n$ matrix M , whose rows are associated with the m elements of the ground set G , whose columns are associated with the n subsets of \mathcal{G} , and whose entries are 1 if the element of the ground set associated with the row is contained in the subset associated with the columns, and 0 otherwise.

SCP is listed among the NP-complete problem class [16], and is therefore considered to be a difficult problem to solve according to the fact that the solution time of any known algorithm cannot be bounded by a polynomial in the size the of the problem, unless of course, $P = NP$. Given their simplicity and generality, SCPs arise naturally in modeling many real-life problems, some interesting and large-sized examples of which can be found, among others, in crew scheduling and allocation ([6, 7, 5]), data mining ([13, 4]), or ([9, 10, 36]).

Therefore, a large effort has been devoted by the research community to find efficient algorithms for its solution. Extensive discussion on algorithms to solve SCP can be found in several surveys, ranging from the 1975 Christofides [11] to the more recent work of Fischetti et al. [8]. An important stream of research on the issue is devoted to approximation results for SCP, where solution algorithms are evaluated in their ability to find a solution whose distance from the optimum is guaranteed with a given probability. Since SCP is NP-hard, numerous heuristic algorithms – mostly of the greedy type – have been developed for its solution, and the best approximation ratio available in polynomial time is $H_d = \sum_{k=1}^d \frac{1}{k}$, i.e. $\Theta(\log d)$, where d is the size of the largest subset (as from [12, 14, 20, 28]), assuming $P \neq NP$ (recall that the approximation ratio of an algorithm is the ratio between the cost of the solution obtained by the algorithm and the cost of an optimal solution).

More recently, a result by Levin [26] provides an approximation ratio of $H_d - \frac{196}{390}$, which improves the previous results on the approximation ratio of the greedy algorithm. For the experimental results and comparison of the performance of many different approximation algorithms of SCP, see e.g. [19, 25].

In 1984 Vercellis [35], guided by previous results [12, 28], studied SCP problems with certain properties, namely defined by a matrix M with random i.i.d. Bernoulli entries. In that paper a class of randomized algorithms which find almost surely a solution whose approximation ratio tends asymptotically to 1 was exhibited, providing the asymptotic cardinality of the optimal solution of SCP problems associated with random matrices in the above sense.

In this paper we prove, via the so-called *probabilistic method in combinatorics* (see [1]), that the asymptotic SCP cardinality value for random matrices with fixed density δ found in [35] is actually an upper bound valid for any matrix with maximum row density δ and any fixed dimension. Moreover, we show how such a-priori bound may be tailored for matrices with uneven row densities, and how the use of more refined approximations in the computations could result in a (sub-leading) improvement of its value.

In addition, we define the class of the (ν, μ) -decomposable 0 – 1 matrices and we show that when the matrix of an SCP belongs to this class, an improved bound can be obtained according to the depth of the decomposition. Such fact indicates an interesting direction for applications, where SCP problems that are not perfectly decomposable may be treated with approximate decomposition algorithms guided by the evaluation of our bound.

The paper is organized as follows: in Section 2 we provide the basic notation that will be used throughout the paper. Section 3 describes the main result and its extensions; Section 4 discusses the refinement of the bound in the case of (ν, μ) -decomposable matrices. Some conclusions are drawn in Section 5.

2 Notation and Previous Results

A formal definition of SCP is given below:

Definition 2.1. Let $G = \{g_1, g_2, \dots, g_m\}$ be a ground set of m elements, and let $\mathcal{G} \subset 2^G$ be a collection of subset of G , $|\mathcal{G}| = n$, where $\cup_{S \in \mathcal{G}} S = G$ and each S has a positive cost c_S . $S_j \in \mathcal{G}$

covers $g_i \in G$ if $g_i \in S_j$. $\mathcal{C} \subset \mathcal{G}$ is said to be a cover of G if $\cup_{S \in \mathcal{C}} S = G$. A minimal cover of G is a cover for which $\sum_{S \in \mathcal{C}} c_S$ is minimum. Given G , \mathcal{G} , and the associated costs $c_S, S \in \mathcal{G}$, the Set Covering Problem SCP amounts to finding a minimal cover of G .

Without loss of generality, we assume that $G = \{1, 2, \dots, m\}$ and $\mathcal{G} = \{S_j, j \in \{1, 2, \dots, n\}\}$.

As said in the introduction, we can describe SCP as the problem of covering the rows of a $m \times n$ matrix M , whose rows are associated with the m elements of the ground set, whose columns are associated with the n subsets of \mathcal{G} , and whose entries are 1 if the element of the ground set associated with the row is contained in the subset associated with the columns, and 0 otherwise. More formally

Definition 2.2. Given a $m \times n$ matrix $M = (m_{ij})$, where

$$m_{ij} = \begin{cases} 1, & \text{if column } j \text{ with associated cost } c_j \text{ covers row } i \\ 0, & \text{otherwise} \end{cases}$$

then, SCP seeks the subset of columns that covers all rows, whose sum of costs is minimal. When the cost $c_i = 1$ for all i , the problem is called SCP with unit costs and the solution is given by the cover of minimal cardinality.

In this paper we deal with SCP with unit costs; in the following all SCPs are assumed to be of that type.

Remark. Observe that if M is a 0–1 matrix describing the set covering problem of a given ground set G with a given collection \mathcal{G} of subsets of G , the any other matrix M' obtained from M by permutations of its rows and/or columns describes the same set-covering problem as the original matrix M .

Let us summarize briefly the results on SCP for random matrices. As remarked in the introduction, the first result, obtained in [35], is related to matrices in which m_{ij} are i.i.d. 0–1 Bernoulli variable with probability δ . This model is known as the *constant density model for SCP*.

Theorem 2.1 (Vercellis). Let C_m be the random variable that represent the optimal cost of random SCP. Suppose that the following two condition are satisfied:

$$C1 : \quad \lim_{m \rightarrow \infty} \frac{n}{\log m} = \infty,$$

$$C2 : \text{ there exist } \alpha > 0 \text{ such that } n \leq m^\alpha.$$

Then the sequence of random variables C_m satisfies

$$\lim_{m \rightarrow \infty} \frac{C_m}{\log m} = \left[\log \frac{1}{1 - \delta} \right]^{-1} \quad a.s. \quad (2.1)$$

Note that, when m and n are asymptotically large, the optimal cost is given by

$$C_m = \frac{\log m}{|\log(1 - \delta)|}$$

with probability 1.

For $i = \{1, 2, \dots, m\}$, let δ_i be the density of 1's of the row i each row (simply called *row density* from now on), i.e.,

$$\delta_i = \frac{1}{n} \sum_{j=1}^n m_{ij}$$

A second model, introduced by Karp [22], assumes that there is an equal number of ones in each row of the matrix M , that is,

$$\delta_i = \frac{1}{n} \sum_{j=1}^n m_{ij} = \delta, \forall i$$

Note that in the Karp model the random variables m_{ij} are not independent for $j \in \{1, 2, \dots, n\}$, but they are indeed independent for $i \in \{1, 2, \dots, m\}$. These models have been studied by Fontanari [15] using statistical mechanics techniques, which are useful in the study of combinatorial optimization problems, see e.g. Mezard, Parisi and Virasoro [30]. The main result of Fontanari's work is that, for the Karp model, the lower bound for the optimal cost is the same obtained by Vercellis in the constant density model.

3 An *a-priori* Bound

Let M be a given 0–1 matrix with m rows and n columns. Solving SCP for M corresponds to find a set $J \subset \{1, 2, \dots, n\}$ of columns of M of minimal cardinality $|J|$ such that for all $i \in \{1, 2, \dots, m\}$

$$\sum_{j \in J} m_{ij} > 0$$

We are in particular interested in a possible a-priori estimate of the minimal cardinality $k = |J|$ of J as a function of the densities δ_i . To get such upper bound we will use the so-called probabilistic method in combinatorics.

The philosophy of the probabilistic method is to prove the existence of combinatorial objects with certain desirable properties (e.g. a proper coloring of the edges of a graph) by showing that these objects have a positive probability to occur in some suitably defined probability space. In particular, the method works as follows. Suppose we are able to define a probability space in which the occurrence of the combinatorial object with the desirable property – the “good event” A – is ensured if a collection of “bad events” $\{B_1, \dots, B_m\}$ is such that none of them occur.

Namely we assume that we are able to define a probability space in which the good event A can be written as

$$A = \bigcap_{i=1}^m \bar{B}_i$$

where \bar{B}_i denote the probabilistic complement of B_i (i.e. \bar{B}_i is the event that B_i does not occur). Suppose then to be able to calculate (or to give an upper bound of) the probability $P(B_i)$ of occurrence for each of the bad events. Then, the probability of the event A is given by

$$P(A) = P\left(\bigcap_{i=1}^m \bar{B}_i\right) = 1 - P\left(\bigcup_{i=1}^m B_i\right)$$

Regardless the structure of dependencies of events B_i we can write

$$P(\cup_{i=1}^m B_i) \leq \sum_{i=1}^m P(B_i) \quad (3.1)$$

Thus the good event A occurs with positive probability if

$$\sum_{i=1}^m P(B_i) < 1 \quad (3.2)$$

We note that inequality 3.2 is the well-known Local Lovász Lemma condition (see, e.g., [1]) when, as it is in our case, each bad event depends on all the others.

This philosophy can be applied to SCP for the fixed matrix M in a quite straightforward way. Indeed, consider a probability space in which the elementary events are the uniformly random choices of a set J with fixed cardinality $|J| = k$ of columns in the matrix M . Define m bad events B_1, \dots, B_m with B_i being the event that $\sum_{j \in J} m_{ij} = 0$. In other words, B_i is the event that the i 'th row is not covered by the columns in the set J . Then the good event A is the event that “every row is covered by at least a column of the set J ” and A clearly occurs if none of the events B_i occur. It is immediate to see that the probability $p_i = P(B_i)$ is such that

$$p_i \leq (1 - \delta_i)^k \quad (3.3)$$

Indeed,

$$p_i = \frac{\binom{n - \delta_i n}{k}}{\binom{n}{k}} = (1 - \delta_i)^k \frac{1 - \frac{1}{n(1 - \delta_i)}}{1 - \frac{1}{n}} \frac{1 - \frac{2}{n(1 - \delta_i)}}{1 - \frac{2}{n}} \dots \frac{1 - \frac{k-1}{n(1 - \delta_i)}}{1 - \frac{k-1}{n}} \leq (1 - \delta_i)^k$$

Now, using the condition (3.2), we have that a covering J of cardinality k exists if

$$\sum_{l \in \{1, 2, \dots, m\}} (1 - \delta_l)^k < 1 \quad (3.4)$$

Hence we have proved the following theorem:

Theorem 3.1. *Given the $m \times n$ matrix M as defined above with density δ_i for the i -th row, it always exists a covering J of cardinality k given by*

$$k = \min\{i \in \{1, \dots, n\} \mid \sum_{l \in \{1, 2, \dots, m\}} (1 - \delta_l)^i < 1\} \quad (3.5)$$

Letting $\delta = \max_i\{\delta_i\}$ be the maximal row density of the matrix M , we get immediately the following corollary

Corollary 3.1. *Given the $m \times n$ matrix M defined above if the density δ_i for the i -th row does not exceed δ , then there exists a covering J of cardinality*

$$k > \frac{\log m}{|\log(1 - \delta)|} \quad (3.6)$$

Remark. One may ask how good are the bounds obtained by Theorem 3.1 and Corollary 3.1. Recalling the result of [35], and in particular formula 2.1 in section 2, we can observe that for matrices in which the only information available is the maximum row density δ the bound (3.6) is optimal in the sense that it is possible to exhibit an example of a matrix M for which the optimal solution of the SCP has the cardinality given by the r.h.s. of (3.6) asymptotically in m, n . Such matrix would, according to [35], belong to the class of random matrices where entry is 1 with probability δ and 0 otherwise. In other words, combining our result with that of [35], we can claim that the random matrices with constant density δ have the worst possible optimal solution for the set covering problem (in the sense of the largest cardinality); i.e., any other matrix M with maximal (or even constant) row-density δ has an optimal solution with cardinality less or equal than that of the random matrix with density δ .

3.1 Further refinements

The above bounds (3.5) and (3.6) can be improved when additional information on the structure of the matrix M , beside the row densities δ_i , is available. This yields anyway into sub-leading corrections to the asymptotic bounds (3.5) and (3.6). The idea is the following. Starting from equation (3.1), we can give a better bound of the quantity $P(\cup_{i=1}^m B_i)$ using the Bonferroni inequality. Indeed, instead of the trivial inequality (3.1), we can write

$$P(\cup_{i=1}^m B_i) \leq \sum_{i=1}^m P(B_i) - \sum_{1 \leq i < j \leq m} P(B_i \cap B_j) + \sum_{1 \leq i < j < k \leq m} P(B_i \cap B_j \cap B_k) \quad (3.7)$$

The two probabilities $P(B_i \cap B_j)$ and $P(B_i \cap B_j \cap B_k)$ can be evaluated as follows. Let $\Gamma_{ij} = \{l \in \{1, \dots, n\} \mid m_{il} = m_{jl} = 1\}$. In other words, the set Γ_{ij} represents the overlap between the two rows i and j . We now define $\gamma_{ij} = \frac{|\Gamma_{ij}|}{n}$. Moreover, let $\Gamma_{ijk} = \{l \in \{1, \dots, n\} \mid m_{il} = m_{jl} = m_{kl} = 1\}$ and define $\gamma_{ijk} = \frac{|\Gamma_{ijk}|}{n}$. Then, we can write

$$P(B_i \cap B_j) = \frac{\binom{n(1-\delta_i-\delta_j+\gamma_{ij})}{k}}{\binom{n}{k}} \quad (3.8)$$

$$P(B_i \cap B_j \cap B_k) = \frac{\binom{n(1-\delta_i-\delta_j-\delta_k+\gamma_{ij}+\gamma_{ik}+\gamma_{jk}-\gamma_{ijk})}{k}}{\binom{n}{k}} \quad (3.9)$$

Then we can identify our upper bound on SCP finding the smallest k such that

$$\begin{aligned} & \sum_{i=1}^m \binom{n(1-\delta_i)}{k} - \sum_{1 \leq i < j \leq m} \binom{n(1-\delta_i-\delta_j+\gamma_{ij})}{k} + \\ & + \sum_{1 \leq i < j < k \leq m} \binom{n(1-\delta_i-\delta_j-\delta_k+\gamma_{ij}+\gamma_{ik}+\gamma_{jk}-\gamma_{ijk})}{k} < \binom{n}{k} \end{aligned} \quad (3.10)$$

The above condition is easy to be checked numerically. Clearly there are choices of γ_{ij} and γ_{ijk} for which the condition above gives an estimate for k which is better than (3.6). To give a flavour, let us consider a random matrix with constant density δ . We have then $\gamma_{ij} = \delta^2$ for all $1 \leq i < j \leq m$ and $\gamma_{ijk} = \delta^3$ for all $1 \leq i < j < k \leq m$. Condition (3.10) becomes, neglecting $o(m)$ terms

$$y - \frac{y^2}{2} + \frac{y^3}{6} < 1$$

where $y = m(1-\delta)^k$. This gives, instead of (3.6), the condition

$$k > \frac{\log m - \log(1.56)}{|\log(1-\delta)|}$$

In the (easy) latter case of the random matrix, the constant can be further improved by considering the intersections up to 5, 7, 9... sets B_i . This gives, in place of 1.56, larger and larger constants, and when we arrive up to the intersections of m sets, for n exponentially large in m , we obtain $k > 0$. This is not surprising, since if the random matrix has a number of column exponentially large, then, almost surely, we have a column j with $m_{ij} = 1 \quad \forall i$.

4 Improvements for partially decomposable problems

The remark after Definition 2.2 suggests that it may be convenient to permute rows and columns of a 0–1 matrix M and to decompose it in blocks, in order to try to improve the bound (3.6). Indeed, the decomposition of a matrix is an alternative representation that allows a particular structure to emerge. Such structures attract the interest of researchers as they may facilitate the solution of certain mathematical problem where the decomposed matrix plays a role. It is well established (see e.g. [3]) that decompositions allow to confine and control particularly “difficult” substructures of the matrix, and moreover allow to parallelize solution algorithms over the substructures – typically, *blocks* – identified by the decomposition. A comprehensive

analysis of the different types of matrix decomposition and the related algorithms is beyond the scope of this paper. Recently, similar problems are discussed in [21] and [2]; the relevant interactions between the block decomposition of binary matrix and several data mining problems are highlighted in [27, 37].

Here we define a class of decomposable 0–1 matrices with maximum row density δ , that we assume to possess an interesting structure, and exploit the extension of the bound (3.6) for SCP whose associated matrix M belongs to this class. The general idea of this class is that the matrix can be decomposed into four block matrices such that the maximum row density of the two block matrices in the main diagonal is larger than δ , while the maximum row density of the remaining off-diagonal two matrices is smaller than δ . Such definition is indeed similar to the *bordered block diagonal form* treated in [3], where its interest for optimization problems is discussed and several decomposition algorithms are referred.

Definition 4.1. *Let M be a 0–1 matrix of dimension $m \times n$ with maximum row density δ . Let $\nu, \mu \in (-1, 1)$. Then M is (ν, μ) -**decomposable** if, after a permutation of its rows and columns, becomes a 0–1 matrix M' formed by the 4 submatrices $M_{11}, M_{12}, M_{21}, M_{22}$ such that:*

$$M' = \begin{pmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{pmatrix}$$

and

- M_{11} has $\frac{m}{2}(1 + \mu)$ rows, $\frac{n}{2}(1 + \nu)$ columns, and maximum row density $\delta_1 > \delta$,
- M_{12} has $\frac{m}{2}(1 + \mu)$ rows, $\frac{n}{2}(1 - \nu)$ columns, and maximum row density $\delta_2 < \delta$,
- M_{21} has $\frac{m}{2}(1 - \mu)$ rows, $\frac{n}{2}(1 + \nu)$ columns, and maximum row density $\delta_3 < \delta$,
- M_{22} has $\frac{m}{2}(1 - \mu)$ rows, $\frac{n}{2}(1 - \nu)$ columns, and maximum row density $\delta_4 > \delta$.

We want to exploit the case of (ν, μ) -decomposable matrices with the following random experiment. Choose uniformly at random k_1 columns of the matrix M in the first $\frac{n}{2}(1 + \nu)$ columns, and k_2 columns in the following $\frac{n}{2}(1 - \nu)$ columns.

Call B_i the (bad) event to have that the $k_1 + k_2$ columns do not cover the row i , with $i = 1, 2, \dots, \frac{m}{2}(1 + \mu)$, and call \tilde{B}_i the (bad) event to have that the $k_1 + k_2$ columns do not cover the row i , with $i = \frac{m}{2}(1 + \mu) + 1, \dots, m$. The probabilities of such bad events are, reasoning as in section 3, bounded by:

$$\begin{aligned} P(B_i) &\leq (1 - \delta_1)^{k_1} (1 - \delta_2)^{k_2} \\ P(\tilde{B}_i) &\leq (1 - \delta_3)^{k_1} (1 - \delta_4)^{k_2} \end{aligned} \tag{4.1}$$

Calling $B = \cup_{i=1}^{\frac{m}{2}(1+\mu)} B_i$ and $\tilde{B} = \cup_{i=\frac{m}{2}(1+\mu)+1}^m \tilde{B}_i$, we are looking for an a-priori estimate of the probability $P(B \cup \tilde{B})$ of the event $B \cup \tilde{B}$ of the form

$$P(B \cup \tilde{B}) = P(B) + P(\tilde{B}) - P(B \cap \tilde{B}) < 1 \tag{4.2}$$

because this relation would imply, as before, the fact that the complementary event $A = \overline{B \cup \tilde{B}}$, which is the (good) event to have all the rows covered by the $k_1 + k_2$ columns chosen by our random experiment, would have a probability strictly positive, and hence it would exist.

The trouble with (4.2) is the fact that it is not easy, in general, to give a non trivial lower bound of the quantity $P(B \cap \tilde{B})$. Here we will use the trivial bound $P(B \cap \tilde{B}) \geq 0$, and we will get rid of such term from the inequality. There are, however, some specific cases in which it is easy to estimate that intersection: for instance, if δ_2 and δ_3 are zero it is easy to see that B and \tilde{B} are independent, and then $P(B \cap \tilde{B}) = P(B) \times P(\tilde{B})$. We will recall this later; for the time being let us write our condition in terms of the following inequality:

$$P(B \cup \tilde{B}) \leq P(B) + P(\tilde{B}) < 1 \quad (4.3)$$

Plugging (4.1) in (4.3), we get that if k_1 and k_2 are such that

$$\frac{m}{2}(1 + \mu)(1 - \delta_1)^{k_1}(1 - \delta_2)^{k_2} + \frac{m}{2}(1 - \mu)(1 - \delta_3)^{k_1}(1 - \delta_4)^{k_2} < 1 \quad (4.4)$$

then the good event $A = (B \cup \tilde{B})$ to have all the rows covered by the $k_1 + k_2$ columns has a positive probability to occur. The condition (4.4) can be separated in the two independent bounds:

$$\begin{aligned} \frac{m}{2}(1 + \mu)(1 - \delta_1)^{k_1}(1 - \delta_2)^{k_2} &< \alpha \\ \frac{m}{2}(1 - \mu)(1 - \delta_3)^{k_1}(1 - \delta_4)^{k_2} &< 1 - \alpha \end{aligned} \quad (4.5)$$

with $\alpha \in (0, 1)$ to be determined in order to optimize globally our bound. The two conditions in (4.5) can be rewritten as

$$\begin{aligned} k_1 |\log(1 - \delta_1)| + k_2 |\log(1 - \delta_2)| &> c_1(\alpha) \\ k_1 |\log(1 - \delta_3)| + k_2 |\log(1 - \delta_4)| &> c_2(\alpha) \end{aligned} \quad (4.6)$$

with

$$\begin{aligned} c_1(\alpha) &= |\log \alpha| + \log \left[\frac{m}{2}(1 + \mu) \right] \\ c_2(\alpha) &= |\log(1 - \alpha)| + \log \left[\frac{m}{2}(1 - \mu) \right] \end{aligned} \quad (4.7)$$

Let us consider the following linear system in k_1 and k_2

$$\begin{cases} k_1 |\log(1 - \delta_1)| + k_2 |\log(1 - \delta_2)| = c_1(\alpha) \\ k_1 |\log(1 - \delta_3)| + k_2 |\log(1 - \delta_4)| = c_2(\alpha) \end{cases} \quad (4.8)$$

The solution of such a system is

$$\begin{aligned} k_1 &= \frac{1}{\Delta} (c_1(\alpha) |\log(1 - \delta_4)| - c_2(\alpha) |\log(1 - \delta_2)|) \\ k_2 &= \frac{1}{\Delta} (c_2(\alpha) |\log(1 - \delta_1)| - c_1(\alpha) |\log(1 - \delta_3)|) \end{aligned} \quad (4.9)$$

with $\Delta = \log(1 - \delta_1) \log(1 - \delta_4) - \log(1 - \delta_2) \log(1 - \delta_3)$.

Now let us find the α that minimizes the value of

$$k_1 + k_2 = \frac{1}{\Delta} (c_1(\alpha)[|\log(1 - \delta_4)| - |\log(1 - \delta_3)|] + c_2(\alpha)[|\log(1 - \delta_1)| - |\log(1 - \delta_2)|]) \quad (4.10)$$

It is easy to see that the unique minimum of $k_1 + k_2$, as α varies in $(0, 1)$ is attained for $\alpha = \bar{\alpha}$, with $\bar{\alpha}$ given by

$$\bar{\alpha} = \frac{|\log(1 - \delta_4)| - |\log(1 - \delta_3)|}{|\log(1 - \delta_4)| - |\log(1 - \delta_3)| + |\log(1 - \delta_1)| - |\log(1 - \delta_2)|} \quad (4.11)$$

Recalling now the explicit expression of $c_1(\alpha)$ and $c_2(\alpha)$ given in (4.7), we can put the value of $\alpha = \bar{\alpha}$ given by (4.11) into $c_1(\alpha), c_2(\alpha)$ appearing in (4.10) and get an a priori upper bound for the cardinality $k = k_1 + k_2$ of the optimal solution in the case of the decomposable matrix.

Hence we have proved the following theorem.

Theorem 4.1. *Let M be a (ν, μ) -decomposable matrix as given in definition 4.1. Then there exists a covering J of cardinality $k_1 + k_2$ given by (4.10) with α given by (4.11).*

In general it is not simple to compare analytically the bound (4.10) (putting of course $\alpha = \bar{\alpha}$ given by (4.11)) with the bound (3.6), but one can check numerically that this estimate tends to improve the previous general estimate. In any case, it is thus well established that the relations among $\delta_1, \delta_2, \delta_3, \delta_4$ may play a role in the design of a solution algorithm. We illustrate this fact by considering two examples in which the expression (4.10) simplifies drastically and yet are representative of possible situations. For both these example we get an explicit improvement of the bound (3.6).

Example 1. Suppose that a proper permutation of the rows and the columns of M results in a perfect block decomposition, where, w.l.o.g.,

$$\delta_2 = \delta_3 = 0, \quad \delta_1 = \frac{2\delta}{1 + \nu}, \quad \delta_4 = \frac{2\delta}{1 - \nu} \quad |\nu| < 1 - 2\delta$$

Then the optimal solution of the SCP defined by matrix M can be obtained by the union of the solutions obtained on M_{11} and M_{22} . Indeed in this case (i.e. $\delta_2 = \delta_3 = 0$) the events B and \tilde{B} are clearly independent. Therefore we can write the condition (4.2) in terms of

$$P(B \cup \tilde{B}) = P(B) + P(\tilde{B}) - P(B)P(\tilde{B}) < 1 \quad (4.12)$$

and this is equivalent to impose separately $P(B) < 1$ and $P(\tilde{B}) < 1$. It follows that in formulas (4.6)-(4.10) the factors $c_1(\alpha)$ and $c_2(\alpha)$ can be replaced by

$$\begin{aligned} c_1 &\doteq c_1(1) = \log \left[\frac{m}{2}(1 + \mu) \right] \\ c_2 &\doteq c_2(1) = \log \left[\frac{m}{2}(1 - \mu) \right] \end{aligned} \quad (4.13)$$

and $k_1 + k_2$ is such that

$$k_1 + k_2 = \frac{\log \left[\frac{m}{2}(1 + \mu) \right]}{|\log(1 - \delta_1)|} + \frac{\log \left[\frac{m}{2}(1 - \mu) \right]}{|\log(1 - \delta_4)|} \quad (4.14)$$

I.e., the resulting bound is exactly the general bound given in the section 3 applied separately to the two factorized problems in the blocks 1 and 4.

Let us now show that

$$\frac{\log \left[\frac{m}{2}(1 + \mu) \right]}{|\log(1 - \delta_1)|} + \frac{\log \left[\frac{m}{2}(1 - \mu) \right]}{|\log(1 - \delta_4)|} < \frac{\log m}{|\log(1 - \delta)|}$$

which is equivalent to show that

$$\log(m) \times \left[\frac{1}{|\log(1 - \delta)|} - \frac{1}{|\log(1 - \delta_1)|} - \frac{1}{|\log(1 - \delta_4)|} \right] - \frac{\log \frac{1+\mu}{2}}{|\log(1 - \delta_1)|} - \frac{\log \frac{1-\mu}{2}}{|\log(1 - \delta_4)|} > 0$$

The last two terms do not depend on m and are always non negative (recall that $|\mu| < 1$ and thus $\log \frac{1 \pm \mu}{2} < 0$). Therefore it is enough to prove the the following inequality:

$$\left[\frac{1}{|\log(1 - \delta)|} - \frac{1}{|\log(1 - \delta_1)|} - \frac{1}{|\log(1 - \delta_4)|} \right] > 0$$

i.e. recalling that $\delta_1 = \frac{2\delta}{1+\nu}$ and $\delta_4 = \frac{2\delta}{1-\nu}$, it is enough to prove

$$\frac{1}{|\log(1 - \delta)|} > \frac{1}{|\log(1 - \frac{2\delta}{1+\nu})|} + \frac{1}{|\log(1 - \frac{2\delta}{1-\nu})|} \quad (4.15)$$

To show that the above inequality is always satisfied, first recall that we always have $|\nu| < 1 - 2\delta$. One can now study the expression on the r.h.s. of 4.15, as a function of $\nu \in (-1 + 2\delta, 1 - 2\delta)$. Let

$$f(\nu) = \frac{1}{|\log(1 - \frac{2\delta}{1+\nu})|} + \frac{1}{|\log(1 - \frac{2\delta}{1-\nu})|}$$

It can be checked that $f(\nu)$ is concave in the interval $\nu \in (-1 + 2\delta, 1 - 2\delta)$ and attains the maximum at $\nu = 0$ where reaches the value

$$f(0) = \frac{2}{|\log(1 - 2\delta)|}$$

and since

$$\frac{1}{|\log(1 - \delta)|} > \frac{2}{|\log(1 - 2\delta)|}$$

the inequality above is true for all $\nu \in (-1 + 2\delta, 1 - 2\delta)$.

Example 2. Suppose there exists a permutation of rows and columns of the $m \times n$ matrix M that identifies 4 block matrices of size $\frac{m}{2} \times \frac{n}{2}$, such that the two blocks on the main diagonal

absorb a large portion of the matrix density, at the expenses of the blocks in the other diagonal. Namely suppose that after such a permutation we have

$$\delta_1 = \delta_4 = 2\delta - \epsilon \quad \delta_2 = \delta_3 = \epsilon \quad \mu = \nu = 0 \quad \epsilon < \delta$$

In this case by (4.11) we get that $\bar{\alpha} = \frac{1}{2}$ and thus, plugging this value in formulas (4.7) we get

$$c_1(1/2) = c_2(1/2) = \log m$$

Therefore, considering that in the present case $\Delta = [\log(1 - 2\delta + \epsilon)]^2 - [\log(1 - \epsilon)]^2$ equation (4.10) becomes

$$k_1 + k_2 = \frac{2 \log m}{|\log(1 - 2\delta + \epsilon)| + |\log(1 - \epsilon)|}$$

Let us prove that, for any $\epsilon \in [0, \delta)$ we have

$$\frac{2 \log m}{|\log(1 - 2\delta + \epsilon)| + |\log(1 - \epsilon)|} < \frac{\log m}{|\log(1 - \delta)|}$$

which is equivalent to the inequality

$$|\log(1 - 2\delta + \epsilon)| + |\log(1 - \epsilon)| > 2|\log(1 - \delta)|$$

i.e.

$$\log[(1 - 2\delta + \epsilon) \times (1 - \epsilon)] < \log[(1 - \delta)^2]$$

Inequality above, by the monotonicity of the logarithm, is true if and only if

$$(1 - 2\delta + \epsilon) \times (1 - \epsilon) < (1 - \delta)^2$$

i.e. if

$$2\delta\epsilon - \epsilon^2 < \delta^2$$

which is always true for all $\epsilon \in [0, \delta)$.

5 Conclusions

The results of this paper are related with the existence of an easy to compute *a-priori* upper bound for the Set Covering problem with unit cost. The bound is obtained by the application of the probabilistic method in combinatorics and extends to a deterministic setting previous asymptotic results. We show several variants of the bound that can be computed by a simple binary search, and analyze some extensions. As a side results, we consider the specialization of this bound when the 0 – 1 matrix that describes the SCP can be almost decomposed into a block diagonal matrix. In the latter case we show how the bound is related with the parameters that define the decomposition and show that, under certain conditions, the decomposition always improves the bound.

Although the results presented are mainly related with theoretical properties of the solution of a specific integer programming problem, we believe that they provide an interesting insight for practical application, given the extremely general and simple nature of the bound; moreover, the results of Section 4 suggest that even non-perfect decompositions may be useful to improve solution methods for the hard combinatorial problems considered in this paper. Such considerations demand further investigations and computational tests that will be addressed in future research.

Acknowledgments

This work has been supported by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Consiglio Nazionale delle Ricerche (CNR), and FAPEMIG (Fundação de Amparo à Pesquisa do Estado de Minas Gerais)-Programa Pesquisador Mineiro.

References

- [1] N. Alon; J. Spencer (2008): *The Probabilistic Method. Third Edition*. New York, Wiley-Interscience.
- [2] E. Bertolazzi; A. Rimoldi (2014): *Fast matrix decomposition in F^2* , Journal of Computational and Applied Mathematics, **260**, 519–532.
- [3] R. Borndorfer; C. E. Ferreira; A. Martin (1998): *Decomposing Matrices into Blocks*, Siam Journal of Optimization, **9**, n. 1, 236–269.
- [4] E. Boros; P.L. Hammer; T. Ibaraki (2005): *Logical Analysis of Data*, In: Encyclopedia of Data Warehousing and Mining, (J. Wang, ed.) Idea Group Reference, 689–692.
- [5] M. Boschetti; V. Maniezzo (2014): *A set covering based metaheuristic for a real-world city logistics problem*, International Transactions in Operational Research, doi: 10.1111/itor.12110.
- [6] E.K. Burke; T. Curtois (2014): *New approaches to nurse rostering benchmark instances*, European Journal of Operational Research **237**, 71–81.
- [7] V. Cacchiani; V.C. Hemmelmayr; F. Tricoire (2014): *A set-covering based heuristic algorithm for the periodic vehicle routing problem*, Discrete Applied Mathematics, **163**, 53–64.
- [8] A. Caprara; P. Toth; M. Fischetti (2000): *Algorithms for the Set Covering Problem*, Annals of Operations Research, **98**, 353–371.
- [9] W.A. Chaovalitwongse; T.Y. Berger-Wolf; B. Dasgupta; M.V. Ashley (2007): *Set covering approach for reconstruction of sibling relationships*, Optimization Methods and Software, **22**, 11–24.
- [10] L. Chen; J. Crampton (2009): *Set Covering Problems in Role-Based Access Control*, Lecture Notes in Computer Science **5789**, 689–704.
- [11] N. Christofides; S. Korman (1975): *A Computational Survey of Methods for the Set Covering Problem*, Management Science, **21**, 591–599.

- [12] V. Chvatal (1979): *A greedy heuristic for the set-covering problem*. Math. Oper. Res. **4**, no. 3, pp. 233–235.
- [13] Y. Crama; P.L. Hammer; T. Ibaraki (1988): *Cause-effect relationships and partially defined Boolean functions*, Annals of Operational Research, **16**, 299–325.
- [14] U. A. Fiege (1998): *Threshold of $\ln n$ for approximating set cover*, Journal of the ACM, **45** (4), 634–652.
- [15] J. F. Fontanari (1996): *A statistical mechanics analysis of the set covering problem*, J. Phys. A: Math. Gen., **9**, 473–483.
- [16] M. R. Garey; D. S. Johnson (1979): *Computers and Intractability: A Guide to the Theory of NP-Completeness*, Freeman and Co.
- [17] J. F. Gimpel (1967): *A Stochastic Approach to the Solution of Large Covering Problems*, IEEE Switching and Automata Theory, 76–83.
- [18] O. Goldschmidt; D. S. Hochbaum; G. Yu (1993): *A modified greedy heuristic for the set covering problem with improved worst case bound*, Information Processing Letters archive, 48–6, Dec. 20, 1993, 305–310.
- [19] T. Grossman; A. Wool (1997): *Computational experience with approximation algorithms for the set covering problem*, European Journal of Operational Research, **101**, 81–92.
- [20] D. S. Johnson (1974): *Approximation algorithms for combinatorial problems*, J. Comput. System Sci., **9**, 256–278.
- [21] G. A. A. Kahou; L. Grigori; M. Masha Sosonkina (2008): *A partitioning algorithm for block-diagonal matrices with overlap*, Parallel Computing, **34**, 332–344.
- [22] R. M. Karp (1976): *The probabilistic analysis of some combinatorial search algorithms*, in Algorithms and Complexity: New Directions and Recent Results, 1–20.
- [23] S. Khot; R. Saket (2008): *Hardness of Minimizing and Learning DNF Expressions*, in Proc. FOCS, pp. 231–240.
- [24] M. Krivelevich (1997): *Approximate set covering in uniform hypergraphs*. J. Algorithms **25** , no. 1, pp. 118–143.
- [25] G. Lan (2007): *An effective and simple heuristic for the set covering problem*, European Journal of Operational Research, **176**, 1387–1403.
- [26] A. Levin (2008): *Approximating the unweighted k -set cover problem: greedy meets local search*, SIAM J. Discrete Math., **231**, 25–264.
- [27] T. Li (2005): *A general model for clustering binary data*, in Proceedings of the 11th ACM SIGKDD int. conf. on Knowledge discovery in Data Mining (KDD '05). ACM, New York, NY, USA, 188–197.
- [28] L. Lovasz (1975): *On the ratio of the optimal integral and fractional covers*. Disc. Math. **13**, pp. 383–390.

- [29] C. Lund; M. Yannakakis (1994): *On the hardness of approximating minimization problems*, J. ACM **31** , no. 5, pp. 960–981.
- [30] M. Mezard; G. Parisi; M. A. Virasoro (1987): *Spin glass theory and beyond*, World Scientific, Singapore.
- [31] M. Okun (2005): *On the approximation of the vertex cover problem in hypergraphs*. Discrete Optimization **2**, no. 1, pp. 101–111.
- [32] R. Raz; M. Safra (2007): *A sub-constant error-probability low-degree test, and a subconstant error-probability PCP characterization of NP*. In Proc. STOC, pp. 475–484.
- [33] R. Saket; M. Sviridenko (2012): *New and Improved Bounds for the Minimum Set Cover Problem*, Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques Lecture Notes in Computer Science, Volume 7408, pp 288–300.
- [34] A. Samorodnitsky; L. Trevisan (2000): *A PCP characterization of NP with optimal amortized query complexity*, in Proc. STOC, pp. 191–199.
- [35] C. Vercellis (1984): *A Probabilistic Analysis of the Set Covering Problem*, Annals of Operations Research **1**, 255–271.
- [36] C.N. Vijeyamurthy; R. Panneerselvam (2010): *Literature review of covering problem in operations management*, International Journal of Services, Economics and Management, **2**, 267–285.
- [37] Z. Zhang; T. Li; C. Ding; X. Zhang (2007): *Binary Matrix Factorization with Applications*, in Proceedings of the 2007 Seventh IEEE International Conference on Data Mining (ICDM '07). IEEE Computer Society, Washington, DC, USA, 391–400.